EJERCICIO PRÁCTICO 13: REGRESIÓN LINEAL SIMPLE Y MÚLTIPLE

CONTEXTO

Si bien hoy en día existen muchas herramientas que facilitan la búsqueda y construcción de un modelo de regresión lineal múltiple (RLM), conseguir un modelo adecuado suele ser un desafío.

El objetivo de este ejercicio es practicar el proceso de creación y evaluación de un modelo de regresión lineal simple (RLS) para predecir una variable numérica y su extensión a multivariado.

OBJETIVOS DE APRENDIZAJE

1. Preparar un conjunto de datos para la construcción de modelos RLS y RLM.
2. Iterar en el proceso de selección de variables, creación y evaluación de modelos RLM, hasta conseguir uno que sea confiable y satisfactorio.

ÉXITO DE LA ACTIVIDAD

El equipo es capaz de encontrar modelos RLS y RLM confiables y de buen desempeño al predecir una variable dependiente.

ACTIVIDADES

Un estudio recolectó medidas anatómicas de 247 hombres y 260 mujeres (Heinz et al., 2003). El estudio incluyó nueve mediciones del esqueleto (ocho diámetros y una profundidad de hueso a hueso) y doce mediciones de grosor (circunferencias) que incluyen el tejido. La siguiente tabla detalla las variables registradas en este estudio:

Columna	Descripción	Unidad
Biacromial.diameter	Diámetro biacromial (a la altura de los hombros)	cm
Biiliac.diameter	Diámetro biiliaco (a la altura de la pelvis)	cm
Bitrochanteric.diameter	Diámetro bitrocantéreo (a la altura de las caderas)	cm
Chest.depth	Profundidad del pecho (entre la espina y el esternón a la altura de los pezones)	cm
Chest.diameter	Diámetro del pecho (a la altura de los pezones)	cm
Elbows.diameter	Suma de los diámetros de los codos	cm
Wrists.diameter	Suma de los diámetros de las muñecas	cm
Knees.diameter	Suma de los diámetros de las rodillas	cm
Ankles.diameter	Suma de los diámetros de los tobillos	cm
Shoulder.Girth	Grosor de los hombros sobre los músculos deltoides	cm
Chest.Girth	Grosor del pecho, sobre tejido mamario en mujeres y a la altura de los pezones en varones	cm
Waist.Girth	Grosor a la altura de la cintura	cm
Navel.Girth	Grosor a la altura del ombligo	cm
Hip.Girth	Grosor a la altura de las caderas	cm

Thigh.Girth	Grosor promedio de ambos muslos bajo el pliegue del glúteo	cm
Bicep.Girth	Grosor promedio de ambos bíceps, brazos flectados	cm
Forearm.Girth	Grosor promedio de ambos antebrazos, brazos extendidos palmas hacia arriba	cm
Knee.Girth	Grosor promedio de ambas rodillas, posición levemente flectada, medición arriba de la rótula	cm
Calf.Maximum.Girth	Grosor promedio de la parte más ancha de ambas pantorrillas	cm
Ankle.Minimum.Girth	Grosor promedio de la parte más delgada de ambos tobillos	cm
Wrist.Minimum.Girth	Grosor promedio de la parte más delgada de ambas muñecas	cm
Age	Edad	Años
Weight	Peso	Kg
Height	Estatura	cm
Gender	Género	1: hombre 0: mujer

Referencia: Heinz, G., Peterson, L. J., Johnson, R. W., & Kerk, C. J. (2003). Exploring relationships in body dimensions. Journal of Statistics Education, 11(2).

1. El equipo copia el enunciado del problema asignado como comentarios de un script R.
2. El equipo lee el enunciado, descarga el archivo de datos (EP13 Datos.csv) desde UVirtual y selecciona las columnas para trabajar de acuerdo a las instrucciones.
3. El equipo construye los modelos solicitados usando la muestra correspondiente.
4. El equipo sube el script con las actividades anteriores comentando en detalle los pasos seguidos.

Fuera del horario de clases, cada equipo debe subir el script realizado UVirtual con el nombre "EP13-respuesta-grupo-i", donde i es el número de grupo asignado. Las respuestas deben subirse antes de las 23:30 del viernes 3 de junio.

PREGUNTA (todos los grupos)

Se pide construir un modelo de regresión lineal múltiple para predecir la variable Peso, de acuerdo con las siguientes instrucciones:

1. Definir la semilla a utilizar, que corresponde a los últimos cuatro dígitos del RUN (sin considerar el dígito verificador) del integrante de menor edad del equipo.
2. Seleccionar una muestra de 50 mujeres (si la semilla es un número par) o 50 hombres (si la semilla es impar).
3. Seleccionar de forma aleatoria ocho posibles variables predictoras.
4. Seleccionar, de las otras variables, una que el equipo considere que podría ser útil para predecir la variable Peso, justificando bien esta selección.
5. Usando el entorno R, construir un modelo de regresión lineal simple con el predictor seleccionado en el paso anterior.

6. Usando herramientas para la exploración de modelos del entorno R, buscar entre dos y cinco predictores de entre las variables seleccionadas al azar en el punto 3, para agregar al modelo de regresión lineal simple obtenido en el paso 5.
7. Evaluar los modelos y "arreglarlos" en caso de que tengan algún problema con las condiciones que deben cumplir.
8. Evaluar el poder predictivo del modelo en datos no utilizados para construirlo (o utilizando validación cruzada).

CRITERIOS DE EVALUACIÓN

Categoría	Nivel de logro	Puntos
Datos	Obtienen una muestra de datos para poder crear y evaluar modelos de regresión lineal, cumpliendo las restricciones indicadas en el enunciado (semilla, tamaño, género, posibles variables predictoras, etc.)	3
Modelo de regresión lineal simple (RLS)	Seleccionan, justificando su utilidad de forma convincente, una variable no elegida anteriormente que utilizan para construir correctamente un modelo de RLS para predecir la variable 'Peso'	3
Selección de variables	Seleccionan un conjunto de variables relevantes para predecir la variable 'Peso', utilizando correctamente gráficos y/o utilidades en paquetes de R para explorarlas, desde el conjunto de ocho variables elegidas aleatoriamente como posibles predictores y respetando las otras restricciones indicadas en el enunciado	4
Modelo de regresión lineal múltiple (RLM)	Construyen correctamente un modelo de RLM para predecir la variable 'Peso' agregando las variables seleccionadas anteriormente al modelo RLS que se tiene	2
Confiabilidad de los modelos	Escriben comentarios y código en R correcto que verifica las condiciones que garantizan que tanto el modelo de RLS como el modelo de RLM obtenidos tienen un buen nivel de ajuste y son generalizables, interpretando explícita y correctamente los resultados obtenidos en cada paso y tomando acciones correctivas apropiadas de ser necesarias o comentando los riesgos asociados	4
Calidad predictiva de los modelos	Escriben código R correcto que evalúa la calidad predictiva tanto del modelo de RLS como del modelo de RLM obtenidos, en datos no utilizados para su construcción, y comparando correctamente los desempeños observados	3
Conclusión	Entregan conclusiones correctas y completas, basadas en las evaluaciones realizadas y el proceso de búsqueda de predictores seguido, respecto del modelo de RLM obtenido	2
Código fuente	El script está completo, ordenado y bien indentado, se comenta paso a paso el procedimiento implementado, logrando un programa que es fácil de seguir y que no requiere cambios para que funcione	2
Ortografía y redacción	El script está bien documentado, escritos con buena ortografía y redacción (<3 errores), usando vocabulario propio de la disciplina y el contexto del problema	2
	TOTAL	**25**
	NOTA	**7,0**